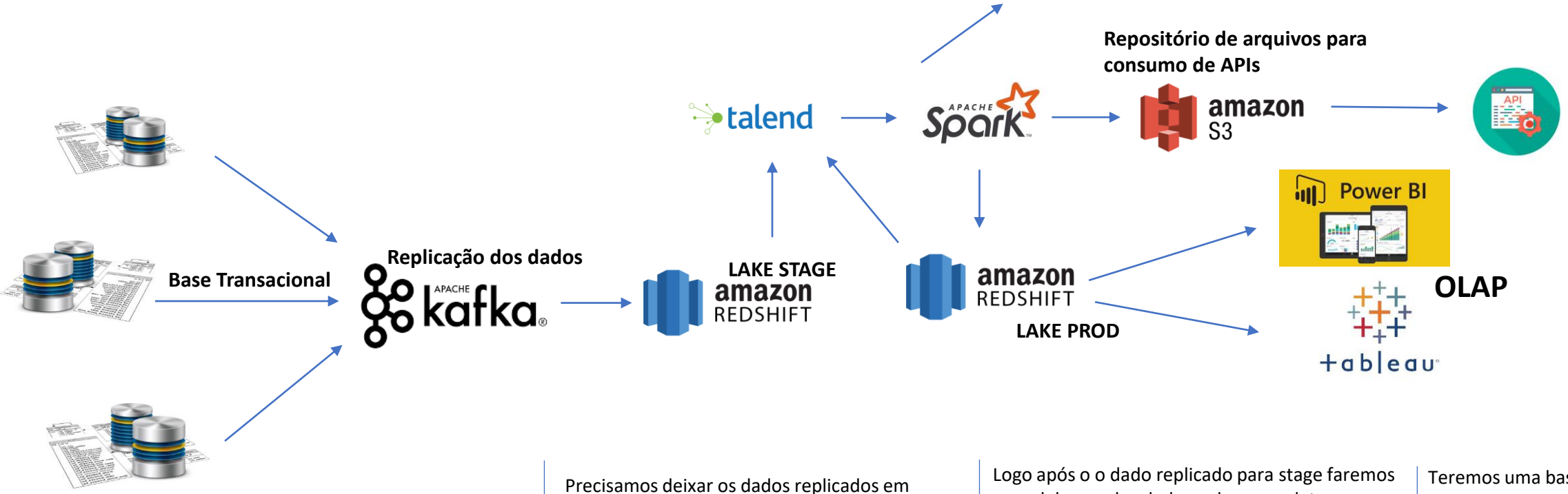
Proposta de Estrutura para Data Lake

Para: Rubens – Itaú

Escrito por: Isaias Trindade Jr

Modelagem dos dados para um Data Lake

Para melhor segurança e controle das alterações iremos armazenar e versionar os códigos gerados em um repositório do GitHub.



Utilizaremos a mensageria Kafka para fazermos a replicação dos dados para uma base na qual podemos fazer as modelagens, via ODBC.

Precisamos deixar os dados replicados em um repositório chamado lake stage, que atenda nossas necessidades.
Para isso escolhi o Redshift, por suportar uma grande massa de dados e escalonamento rápido.

Logo após o o dado replicado para stage faremos a modelagem dos dados e do nosso data warehouse onde ficaram os datamarts separados por business unity.
Essa modelagem será feita através do Talend, exportaremos o código JAVA(job) para o spark executar o trabalho. Isso nos trás mais agilidade na modelagem e facilita o desenvolvimento com alta performance na execução.

Teremos uma base do bucket S3 na qual iremos disponibilizar arquivos para serem consumidos por APIs ou cientistas de dados (eles também podem conectar nos DMs que estarão no Data Lake, porem um repositório para arquivo facilita o consumo dos APIs.

Justificativa geral e ferramentas utilizadas

Minha escolha por utilizar o ambiente da AWS foi por conta de ser um ambiente mais flexível e de custo menor em comparação com a concorrência, sem contar as inúmeras ferramentas que a amazona nos oferece

- Amazon Redshift

O Amazon Redshift é um banco de dados colunar, otimizado para processar grande volume de dados e utilizado na nossa arquitetura para modelagem de dados analíticos e dimensionais OLAP.
Pontos fortes: Grande ganho de velocidade nas consultas, facilidade de uso e acessibilidade, escalonamento rápido, mantém os custos baixos e ferramentas de segurança robustas.

- Amazon S3

Mais conhecido como Bucket s3, o Amazon Simple Storage Service, é um serviço de file System, que fornece amarzenamento de objetos (arquivos) através de uma interface web, acessado via shell script ou por meio da console AWS.

Pontos fortes: Armazenamento de dados e arquivos, download de dados e arquivos, criação de permissões.

- Talend Data Integration

O Talend Data Integration é um conjunto de ferramentas da familia ETL (Extração,Transformação e Carga) para integração de dados. Fornece meios para integrar e processar todos os seus dados com um designer visual fácil de usar.

Pontos fortes: Geração de código JAVA, diversos conectores opensuse , desenvolvimento prático e simplificado.

- GitHub

GitHub é um repositório de versionamento para códigos.

Pontos fortes: Maior segurança na alteração e versionamento dos códigos, grande comunidade e velocidade para pull e push.

- Kafka

Kafka é uma plataforma distribuída de transmissão de dados ele armazena mensagens e usar os tópicos como tabelas, permitindo fazer queries para aquisição de dados.

Pontos fortes: Persistência da mensagem, Balanceamento de carga.

- Spark

O Apache Spark é uma ferramenta Big Data que tem o objetivo de processar grandes volume de dados de forma paralela e distribuída.

Pontos fortes: Velocidade de processamento, execução distribuída.

Agradeço sua atenção !

Obrigado pela oportunidade